

Les principaux formats de fichiers bioinformatiques

Sarah Maman

Sans doute le format de fichier **le plus répandu** car très simple et l'un des plus pratiques.

>gi|22777494|dbj|BAC13766.1| glutamate dehydrogenase [Oceanobacillus iheyensis]

```
MVADKAADSSNVNQENMDVLNTTQTIKSAIDKLGYPEEVFELLKEPMRILTVRIPVRMDDGNVKVFTGY
RAQHNDVAVGPTKGGIRFHPNVTETEVKALSIWMSLKSGIVDLPYGGAKGGIICDPREMSFRELEALSRGY
VRAVSQIVGPTKDIPAPDVFTNSQIMAWMMDEYSKIDFNPNPGFITGKPIVLGGSHGRESATAKGVTVL
NEAAKKKGIDIKGARVVIQGFNGSFLAKFLHDAGAKVVAISDAYGALYDPEGLDIDYLLDRRDSFGTV
TKLFNNTISNDALFELDCDIIVPAAVENQITRENAHNIKASIVVEAANGPTTMEATKILTERDILIVPDV
LASAGGVTVSYFEWVQNNQGFYWSEEEIDNKLHEIMIKSFNNIYNMSKTRRIDMRLAAYMVGVRKMAEAS
```

Un fichier au format FASTA peut contenir plusieurs séquences. Chaque séquence (écrite sous forme de lignes de 80 caractères maximum), est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">".

gi|22777494 : l'identifiant gi ("GenInfo Identifier") est le numéro d'identification d'une séquence (acides aminés ou nucléotides). Si une séquence est modifiée, un nouveau numéro de GI est attribué.

Format FASTA

Fasta: text file

Most basic file format to represent nucleotide or amino-acid sequences.

Each sequence is represented by:

- A single description line (≤ 80 characters)
- The sequence (≈ 70 -80 characters per line, one or several lines)

```
>Protein1 Description of protein 1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGK
LVSVKVSDDFTTAAMRPSYLSYEDLDMTFVENEFYKALVAELEKENEER
>DNA1 Description of dna segment 1
AACTCTCGCGTAGCTCAGAGAAGAGCTTGATCGATCGTGCTGCTGCTA
CCGCTAGTAGCTGTAGATCGTGCTAGTCAGCATCGATGCTAGCTAGCT
```

Description line

The sequence

> Start of the description line

DNA1 Sequence ID and a space

Description Description of the sequence



4 types de formats de fichiers sont couramment utilisés :

FASTQ : format basé sur du texte pour stocker une séquence biologique (généralement la séquence nucléotidique) et des scores de qualité liés à cette séquence (les 2 sont codés par des caractères ASCII sur plusieurs lignes - exemple : la ligne 1 commence avec le caractère @). C'est le fichier de données brutes issues du séquenceur.

SAM ("Sequence Alignment/Map") : format basé sur du texte délimité avec une section en-tête (facultative) et une section alignement. BAM : codage binaire du fichier SAM correspondant.

GTF ("Gene Transfer Format") : format basé sur du texte délimité par des tabulations et des champs. Ce format est utilisé par beaucoup de logiciels pour décrire la structure des transcrits (introns, exons, sites de démarrage, UTR, ...) et le lien entre les transcrits et le gène auquel ils sont associés.

BAM ("Binary Alignment/Map") : format compressé au format de compression BGZF. L'objectif de BGZF est de fournir une bonne compression tout en permettant un accès aléatoire efficace au fichier BAM pour des requêtes indexées.



C'est un format basé sur du texte pour stocker à la fois une séquence biologique (séquence nucléotidique habituellement) et ses scores de qualité.

Une valeur de qualité Q est un nombre entier qui traduit la probabilité que l'appel de la base correspondante est incorrect.

Fichier ci-dessous : première ligne = nom de la séquence après le symbole @ (et, éventuellement, la description) / deuxième ligne = la séquence / quatrième ligne = scores de qualité codés sous forme de lettres.

@sequence 1

ATCGATCAAATAGTCCATTTACAGTTTGGATTTGGGGTCACAGTTTAAGCAGTTTCAACT

+

!''*((((**+))%%%++)) (%%%) .1***-+'') **55CCF>>>>>CCCCC

- 1 séquence = 4 lignes dans le fichier

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*((( (**+))%%%++) ) (%%%) .1***-+'' ) **55CCF>>>>>CCCCCCC65
```

- 1 ère ligne = identifiant de la séquence

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	x'-coordinate of the cluster within the tile
197393	y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (<i>paired-end or mate-pair reads only</i>)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Format FASTQ (suite)

- 4ème ligne = Qualité

```
!''*(((***+))%+%+)(%+%+).1***-+*')**55CCF>>>>>CCCCCCC65
```

- Appelée aussi Phred quality score (Sanger format)

$Q_{\text{sanger}} = -10 \log_{10} p$ Probabilité qu'une base soit incorrecte

- A more compact format to store sequence and qualities
- Can be gzip-ped and used as such by some programs

@SEQ_ID	@ and sequence ID
GATTGTGGGGTTCAAAGCAGTATCGA	Sequence
+	+ (and the sequence ID)
!''*(((***+))%+%+)(%+%+).1***-+*')**55CCF>>>>>CCCCCCC65	The quality score

SAM : Sequence Alignment/ Map format

Les formats SAM et BAM sont les formats standards d'alignement

- To store large nucleotide sequence alignments.
- Represents the alignment of sequences (reads) to a reference sequence (genome):
 - Simple to read and parse (text, tab-delimited)
 - Flexible (possibility to add custom fields)
 - Compact in file size
 - Can store paired-end information

Headers

Sequences

Read

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTACATAAAGCATACTC *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TACCC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

A BAM file (.bam) is the binary version of a SAM file.

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQUENCE on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)



BAM = SAM compressé – Version binaire

--> version compact et indexée pour représenter une séquence de nucléotides alignés.

Indexed BAM : *.bam.bai

Les outils de post process et les visualisateurs utilisent le format BAM pour éviter d'extraire toutes les informations. Par conséquent, le disply du fichier est beaucoup plus rapide vu que seules certaines parties sont accessible au servuer de visualisation.

Voici les étapes bioinformatiques utiles pour générer un BAM :

- 1 – La plateforme de séquençage vous fournit les séquences sous forme de fichiers FASTQ
- 2 – Il est possible de vérifier la qualité de vos séquences avec l'outil FASTQC report.
- 3 – N'hésitez pas à «éliminer les séquences de trop mauvaise qualité --> cf sig-learning.
- 4 – Mapper le FASTQ sur un génome de référence indexé à l'aide de BWA (par exemple).
- 5 – L'indexation du génome de référence est réalisée au niveau du cluster de calcul pour un partage des références.
- 6 – Les résultats du mapping sont obtenus sous forme d'un fichier SAM.
- 7 – La suite d'outils bioionformatiques samtools vous permet ensuite de convertir vos SAM en BAM, et de visualiser le BAM obtenu :

```
samtools view -S -b -o my.bam my.sam
```

- 8 – Le BAM n'est utilisable que si ce dernier est trié et indexé : `samtools sort my.bam my.sorted` puis `samtools index my.sorted.bam`



Il s'agit de formats de Localisation/Annotation/Visualisation

Les 5 formats sont le plus couramment utilisés sont :

BED

GFF --> GTF (dérivé du GFF)

WIG

BEDGRAPH

Ces fichiers ont 1 ligne par zone.

Format BED

BED pour Browser Extensible Data

3 champs obligatoires ; chr, start, stop.

9 autres champs sont optionnels et peuvent contenir :

- Le brin (forward ou reverse)
- Le nom de l'intervalle
- De l'information sur l'intervalle (annotation)

chr1	213941196	213942363
chr1	213942363	213943530
chr1	213943530	213944697
chr2	158364697	158365864
chr2	158365864	158367031
chr3	127477031	127478198
chr3	127478198	127479365
chr3	127479365	127480532
chr3	127480532	127481699

Format GFF

GFF pour General Feature Format

Format utilisé pour localiser et décrire toute zone caractéristique d'un génome (ex : un exon)

Contient 8 champs : Nom, Source, Type, Début d'intervalle, Fin d'intervalle, Score, Brin, Cadre

```
SEQ1 EMBL atg 103 105 . + 0
SEQ1 EMBL exon 103 172 . + 0
SEQ1 EMBL splice5 172 173 . + .
SEQ1 netgene splice5 172 173 0.94 + .
SEQ1 genie sp5-20 163 182 2.3 + .
SEQ1 genie sp5-10 168 177 2.1 + .
SEQ2 grail ATG 17 19 2.1 - 0
```

Format GTF

GTF pour Gene Transfert Format, dérivé du GFF

Contient les mêmes champs + 2 pour l'annotation

```
381 Twinscan CDS 380 401 . + 0 gene_id "001"; transcript_id "001.1";
381 Twinscan CDS 501 650 . + 2 gene_id "001"; transcript_id "001.1";
381 Twinscan CDS 700 707 . + 2 gene_id "001"; transcript_id "001.1";
381 Twinscan start_codon 380 382 . + 0 gene_id "001"; transcript_id "001.1";
381 Twinscan stop_codon 708 710 . + 0 gene_id "001"; transcript_id "001.1";
```



4. Formats « Variant Calling »

2 formats sont couramment utilisés : Format Pileup (format spécifique de l'outil samtools mpileup, moins utilisé maintenant), Format VCF.

Le format VCF est le format par défaut d'un grand nombre de SNP caller dont GATK.

(a) VCF example

meta info starting with '##'

Header	##fileformat=VCFv4.1										
	##fileDate=20110413										
	##source=VCFtools										
	##reference=file:///refs/human_NCBI36.fasta										
	##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">										
	##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">										
	##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">										
	##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">										
	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">										
	##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">										
	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">										
	##ALT=<ID=DEL,Description="Deletion">										
	##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">										
	##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">										
Body	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
	1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
	1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0 1	2/2
	1	5	rs12	A	G	67	PASS	.	GT:DP	1 0:16	2/2:20
	X	100	.	T		.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

header

body

Format VCF

VCF = standard for storing sequence variation

<http://samtools.github.io/hts-specs/VCFv4.1.pdf>

Difficile / impossible de travailler avec les données NGS --> FASTQ dans Word

--> Lecture des fichiers impossible dans les logiciels de bureautique couramment utilisés

Difficile sur une station locale (manque de ressources) :

- * 1 alignement = 4 processeurs + 15 gb Ram (à multiplier par le nombre d'échantillons)
- * Espace de stockage nécessaire important
- * Gestion des sauvegardes

--> Serveur d'application connecté sur cluster de calcul et baie de stockage

1 - Solution commerciales (CLC Bio, NextGene, ...etc)

2 - Galaxy ...

Merci pour votre écoute

Sources et références :

http://biow.sb-roscoff.fr/ecole_bioinfo/training_material/data_formats/file_formats_Alban_Lermine_Olivier_Inizan_2013-11.pdf

Joe Fass <jnfass@ucdavis.edu> and his « Next Generation Sequence Alignment » slides

The Sequence Alignment/Map format and SAM tools. Li et al. 2009 Bioinformatics 25 2078-2079

The variant call format and VCFtools. Danecek et al. 2011 Bioinformatics 27 2156-2518.